# Text Classification Using Novel "*Anti*-Bayesian" Techniques

B. John Oommen[1(✉)], Richard Khoury[2], and Aron Schmidt[2]

[1] School of Computer Science, Carleton University, Ottawa K1S 5B6, Canada
oommen@scs.carleton.ca

[2] Department of Software Engineering, Lakehead University, Thunder Bay P7B 5E1, Canada
{rkhoury,aschmid1}@lakeheadu.ca

**Abstract.** This paper presents a non-traditional "*Anti*-Bayesian" solution for the traditional Text Classification (TC) problem. Historically, all the recorded TC schemes work using the fundamental paradigm that once the statistical features are inferred from the syntactic/semantic indicators, the classifiers themselves are the well-established statistical ones. In this paper, we shall demonstrate that by virtue of the skewed distributions of the features, one could advantageously work with information latent in certain "non-central" quantiles (i.e., those distant from the mean) of the distributions. We, indeed, demonstrate that such classifiers exist and are attainable, and show that the design and implementation of such schemes work with the recently-introduced paradigm of Quantile Statistics (QS)-based classifiers. These classifiers, referred to as Classification by Moments of Quantile Statistics (CMQS), are essentially "Anti"-Bayesian in their *modus operandi*. To achieve our goal, in this paper we demonstrate the power and potential of CMQS to describe the *very* high-dimensional TC-related vector spaces in terms of a limited number of "outlier-based" statistics. Thereafter, the PR task in classification invokes the CMQS classifier for the underlying multi-class problem by using a linear number of pair-wise CMQS-based classifiers. By a rigorous testing on the standard 20-Newsgroups corpus we show that CMQS-based TC attains accuracy that is comparable to the best-reported classifiers. We also propose the potential of fusing the results of a CMQS-based method with those obtained from a traditional scheme.

**Keywords:** Text classification · Quantile statistics (QS) · Classification by the moments of QS (CMQS)

## 1 Introduction

This paper presents a non-traditional and totally novel solution to the problem of Text Classification (TC). TC is the challenge of associating a given unknown

The authors are grateful for the partial support provided by NSERC, the Natural Sciences and Engineering Research Council of Canada.

B. John Oommen—*Chancellor's Professor*; *Fellow: IEEE* and *Fellow: IAPR*. This author is also an *Adjunct Professor* with the University of Agder in Grimstad, Norway.


M. Núñez et al. (Eds.): ICCCI 2015, Part I, LNAI 9329, pp. 1–15, 2015.
DOI: 10.1007/978-3-319-24069-5_1

text document with a category selected from a predefined set of categories (or classes) based on its content. As opposed to this, statistical Pattern recognition (PR) is the process by which unknown *statistical* feature vectors are categorized into groups or classes based on their *statistical* components [3]. The field of statistical PR has been so well developed that it is not necessary for us to survey the field here. Suffice it to mention that all the recorded TC schemes work using the fundamental paradigm that once the statistical features are inferred from the *syntactic or semantic* indicators, the classifiers themselves are the well-established *statistical*, neural or fuzzy ones such as the Bayesian, Naïve Bayesian, Linear Discriminant, the SVM, the Back-propagation etc.

The TC problem has been studied since the 1960's [13], but it has taken a special importance in recent years as the sheer amount of text available has increased super-exponentially – thanks to the internet, text-based communications such as e-mail, tweets and text messages, and the numerous book-digitization projects that have been undertaken by the various publishing houses. Over the decades, many approaches[1] have been proposed to accomplish this goal. When it concerns classification and PR, the TC problem is particularly interesting both from an academic and a research perspective. This is because, whereas the features in TC are derived from *syntactic or semantic* indicators, the classification, in and of itself, is based on *statistical*, neural or fuzzy strategies.

The goal of this paper is to show that we can achieve TC using "Anti"-Bayesian quantile statistics-based classifiers which only use information contained in, let us say, non-central quantiles (which are sometimes outliers) of the distributions, and that it can do this by operating with a philosophy that is totally contrary to the acclaimed Bayesian paradigm. Indeed, the fact that such a classification can be achieved is, strictly speaking, not easy to fathom.

To motivate this paper and to place its contribution the right context, we present the following simple example. Consider the problem of distinguishing a document that belongs to one of two classes, namely, *Sports* or *Business*. It is obvious that one can trivially distinguish them if we merely considered those words which occurred frequently in one class and not the other, for example, "football" and "basketball" *versus* "dollars" and "euros". Our hypothesis is that it is not *merely* these truly "distinguishing" words that possess "discriminating" capabilities. We intend to demonstrate that there are "outliers" quantiles of the words which occur in both categories, and which also can be used to achieve the classification. Hopefully, this would be both a pioneering and remarkable result.

It should, first of all, be highlighted that we do not intend to obtain a classification that *surpasses* the behavior of the scheme that involves a Bayesian strategy invoking the truly "distinguishing" words. Attempting to do this would be tantamount to accomplishing the impossible, because the Bayesian approach

[1] Due to space limitations, it is impossible to survey the field of TC here. The unabridged version of this paper [8] contains a more detailed survey of the field and includes the preliminaries of the Vector Space model, the Bag-of-Words (BoW), the Term Frequency (TF), the Term Frequency-Inverse Document Frequency (TFIDF) weighting schemes, and the Cosine Similarity metric etc. [11,12].

maximizes the *a posteriori* probability and it thus yields the optimal hallmark classifier. What we endeavor to do is to show that if we use the above-mentioned non-central quantiles and work within an "Anti"-Bayesian paradigm using only *these* quantile statistics, we can obtain accuracies comparable to this optimal hallmark! Indeed, we demonstrate that a near-optimal solution can be obtained by invoking counter-intuitive features *when they are coupled with* a counter-intuitive PR paradigm.

As a backdrop, we note that the basic concept of traditional *parametric* classification is to model the classes based on the assumptions related to the underlying class *distributions*, and this has been historically accomplished by performing a learning phase in which the moments, i.e, the mean, variance etc. of the respective classes are evaluated. However, there have been some families of indicators (or distinguishing quantifiers) that were until recently, noticeably, *uninvestigated* in the PR literature. Specifically, we refer to the use of phenomena that have utilized the properties of the *Quantile Statistics* (QS) of the distributions. This has led to the "Anti"-Bayesian methodology alluded to.

### 1.1 Contributions of this Paper

The novel contributions of this paper are:

– To demonstrate that text and document classification can be achieved using an "Anti"-Bayesian methodology;
– To show that this "Anti"-Bayesian PR can be achieved using syntactic information that that has not been used in the literature before, namely the information contained in the symmetric quantiles of the distributions, and which are traditionally considered to be "outlier"-based;
– To show that the results of our "Anti"-Bayesian PR is not highly correlated with the results of any of the traditional TC schemes, implying that one can use it in conjunction with a traditional TC scheme for an ensemble-based classifier;
– To suggest that a strategy that incorporates the fusion of the features and methodology proposed here and the distinct ones from the state-of-the-art has great potential. This is an avenue that we will explore in future research.

As in the case of the quantile-based PR results, to the best of our knowledge, the pioneering nature and novelty of these TC results hold true.

## 2 Background: Traditional Text Classifiers

Apart from the methods presented above, many authors have also looked at ways of enhancing the document and class representation by including not only words but also bigrams, trigrams, and $n$-grams in order to capture common multi-word expressions used in the text [4]. Likewise, character $n$-grams can be used to capture more subtle class distinctions, such as the distinctive styles of different authors for authorship classification. While these approaches have, so far, considered ways to enrich the representation of the text in the word vector, other authors have attempted to augment the text itself by adding extra information into it, such as synonyms of the words taken from a thesaurus, be it a specialized custom-made one for a project such as the affective-word thesaurus built in [7], or, more commonly, the more general-purpose linguistic ontology, *WordNet* [5].

Adding another generalization step, it is increasingly common to enrich the text not only with synonymous words but also with synonymous *concepts*, taken from domain-specific ontologies [17] or from Wikipedia [1]. Meanwhile, in an opposing research direction, some authors prefer to simplify the text and its representation by reducing the number of words in the vectors, typically by grouping synonymous words together using a Latent Semantic Analysis (LSA) system or by eliminating words that contribute little to differentiating classes as indicated by a Principal Component Analysis (PCA) [6]. Other authors have looked at improving classification by mathematically transforming the sparse and noisy category word space into a more dense and meaningful space. A popular approach in this family involves Singular Value Decomposition (SVD), a projection method in which the vectors of co-occurring words would project in similar orientations, while words that occur in different categories would be projected in different orientations.This is often done before applying LSA or PCA modules to improve their accuracy. Likewise, authors can transform the word-count space to a probabilistic space that represents the likelihood of observing a word in a document of a given category. This is then used to build a probabilistic classifier, such as the popular Naïve-Bayes' classifier [10], to classify the text into the most probable category given the words it contains.

An underlying assumption shared by all the approaches presented above is that one can classify documents by comparing them to a representation of what an average or typical document of the category should look like. This is immediately evident with the BOW approach, where the category vector is built from average word counts obtained from a set of representative documents, and then compared to the set of representative documents of other categories to compute the corresponding similarity metric. Likewise, the probabilities in the Naïve-Bayes' classifier and other probability-based classifiers are built from a corpus of typical documents and represent a general rule for the category, with the underlying assumption that the more a specific document differs from this general rule, the less probable it is that it belongs to the category. The addition of information from a linguistic resource such as a thesaurus or an ontology is also based on this assumption, in two ways. First, the act itself is meant to add words and concepts that are missing from the specific document and thus make it more like a typical document of the category. Secondly, the development of these resources is meant to capture general-case rules of language and knowledge, such as "these words are typically used synonymously" or "these concepts are usually seen as being related to each other."

The method we propose in this paper is meant to break away from this assumption, and to explore the question of whether there is information usable for classification outside of the norm, at "the edges (or fringes) of the word distributions", which has been ignored, so far, in the literature.

## 3 CMQS-Based Text Classifiers

### 3.1 How Uni-dimensional"Anti"-Bayesian Classification Works

We shall first describe how uni-dimensional "Anti"-Bayesian classification works, and then proceed to explain how it can be applied to TC, which, by definition, involves PR in a highly multi-dimensional feature space. Classification by the Moments of Quantile Statistics[2], (CMQS) is the PR paradigm which utilizes QS in a pioneering manner to achieve optimal (or near-optimal) accuracies for various classification problems. Rather than work with "traditional" statistics (or even sufficient statistics), the authors of [14] showed that the set of *distant* quantile statistics of a distribution do, indeed, have discriminatory capabilities. Thus, as a *prima facie* case, they demonstrated how a generic classifier could be developed for any uni-dimensional distribution. Then, to be more specific, they designed the classification methodology for the Uniform distribution, using which the analogous classifiers for other symmetric distributions were subsequently created. The results obtained were for symmetric distributions, and the classification accuracy of the CMQS classifier exactly attained the optimal Bayes' bound. In cases where the symmetrtic QS values crossed each other, one invokes a *dual* classifier to attain the same accuracy.

Unlike the traditional methods used in PR, one must emphasize the fascinating aspect that CMQS is essentially "Anti"-Bayesian in its nature. Indeed, in CMQS, the classification is performed in a counter-intuitive manner i.e., by comparing the testing sample to a few samples *distant* from the mean, as opposed to the Bayesian approach in which comparisons are made, using the Euclidean or a Mahalonibis-like metric, to *central* points of the distributions. Thus, opposed to a Bayesian philosophy, in CMQS, the points against which the comparisons are made are located at the positions where the Cumulative Distribution Function (CDF) attains the percentile/quantile values of $\frac{2}{3}$ and $\frac{1}{3}$, or more generally, where the CDF attains the percentile/quantile values of $\frac{n-k+1}{n+1}$ and $\frac{k}{n+1}$.

In [9], the authors built on the results from [14] and considered various symmetric and *asymmetric* uni-dimensional distributions within the exponential family such as the Rayleigh, Gamma, and Beta distributions. They again proved that CMQS had an accuracy that attained the Bayes' bound for symmetric distributions, and that it was very close to the optimal for asymmetric distributions.

[2] The authors of [14], [9] and [15] (cited in their chronological order) had initially proposed their theoretical and experimental results as being based on the *Order*-Statistics of the distributions. This was later corrected in [16], where they showed that their results were, rather, based on their *Quantile* Statistics.

### 3.2 TC: A Multi-dimensional "Anti"-Bayesian Problem

Any problem that deals with TC must operate in a space that is very high dimensional primarily the because cardinality of the BOW can be very large. This, in and of itself, complicates the QS-based paradigm. Indeed, since we are speaking about the quantile statistics of a distribution, it implicitly and explicitly assumes that the points can be *ordered*. Consequently, the multi-dimensional generalization of CMQS, theoretically and with regard to implementation, is particularly non-trivial because there is no well-established method for achieving the ordering of multi-dimensional data specified in terms of its uni-dimensional components.

To clarify this, consider two patterns, $\mathbf{x}_1 = [x_{11}, x_{12}]^T = [2,3]^T$ and $\mathbf{x}_2 = [x_{21}, x_{22}]^T = [1,4]^T$. If we only considered the first dimension, $x_{21}$ would be the first QS since $x_{11} > x_{21}$. However, if we observe the second component of the patterns, we can see that $x_{12}$ would be the first QS. It is thus, clearly, not possible to obtain the ordering of the *vectorial* representation of the patterns based on their individual components, which is the fundamental issue to be resolved before the problem can be tackled in any satisfactory manner for multi-dimensional features. One can only imagine how much more complex this issue is in the TC domain – when the number of elements in the BOW is of the order of hundreds or even thousands.

To resolve this, multi-dimensional CQMS operates with a paradigm that is analogous to a Naïve-Bayes' approach, although it, really, is of an *Anti*-Naïve-Bayes' paradigm. Using such a *Anti*-Naïve-Bayes' approach, one can design and implement a CMQS-based classifier. The details of this design and implementation for two and multi-dimensions (and the associated conclusive experimental results) have been given in [15]. Indeed, on a deeper examination of these results, one will appreciate the fact that the higher-dimensional results for the various distributions do not necessarily follow as a consequence of the lower uni-dimensional results. They hold by virtue of the factorizability of the multi-dimensional density functions that follow the *Anti*-Naïve-Bayes' paradigm, and the fact that the $d$-dimensional QS-based statistics are concurrently used for the classification in every dimension.

### 3.3 Design and Implementation: "Anti"-Bayesian TC Solution

**"Anti"-Bayesian TC Solution: The Features.** Each class is represented by two BOW vectors, one for each CMQS point used. For each class, we compute the frequency distribution of each word in each document in that class, and generate a frequency histogram for that word. While the traditional BOW approach would then pick the average value of this histogram, our method computes the area of the histogram and determines the two symmetric QS points. Thus, for example, if we are considering the $\frac{2}{7}$ and $\frac{5}{7}$ QS points of the two distributions, we would pick the word frequencies that encompass the $\frac{2}{7}$ and $\frac{5}{7}$ of the histogram area respectively. The reader must observe the salient characteristic of this strategy: By working with such a methodology, for each word in the BOW, we represent

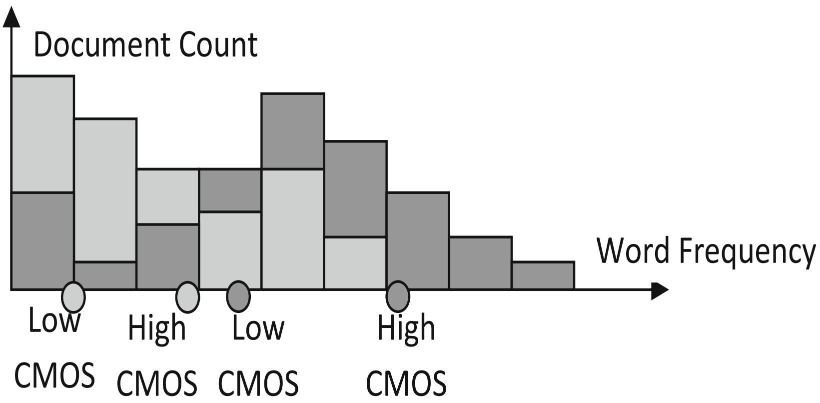


**Fig. 1.** Example of the QS-based features extracted from the histogram of a lower class (light grey) and of a higher class (dark grey), and the corresponding lower and higher CMQS points of each class.

the class by two of its non-central cases, rather than its average/median sample. This renders the strategy to be "Anti"-Bayesian!

For further clarity, we refer the reader to Figure 1. For any word, the histograms of the two classes are depicted in light grey for the lower class, and in dark grey for the higher class. The QS-based features for the classes are then extracted from the histograms as clarified in the figure.

**"Anti"-Bayesian TC Solution: The Multi-Class TC Classifier.** Let us assume that the PR problem involves $C$ classes. Since the "Anti"-Bayesian technique has been extensively studied for two-class problems, our newly-proposed multi-class TC classifier operates by invoking a sequence of $C-1$ pairwise classifiers. More explicitly, whenever a document for testing is presented, the system invokes a classifier that involves a pair of classes from which it determines a winning class. This winning class is then compared to another class until all the classes have been considered. The final winning class is the overall best and is the one to which the testing document is assigned.

**"Anti"-Bayesian TC Solution: Testing.** To classify an unknown document, we compute the cosine similarity between it and the features representing pairs of classes. This is done as follows: For each word, we mark one of the two groups as the high-group and the other as the low-group based on the word's frequency in the documents of each class, and we take the high CMQS point of the low-group and the low CMQS point of the high-group, as illustrated in Figure 1. We build the two class vectors from these CMQS points, and we compute the cosine similarity [8] between the document to classify each class vector.

The most similar class is retained and the least similar one is discarded and replaced by one of the other classes to be considered, and the test is run again, until all the classes have been exhausted. The final class will be the most similar one, and the one that the document is classified into.

# 4 Experimental Set-Up

## 4.1 The Data Sets

For our experiments, we used the 20-Newsgroups corpus, a standard corpus in the literature pertaining to Natural Language Processing. This corpus contains 1,000 postings collected from the 20 different Usenet groups, each associated with a distinct topic, as listed in Table 1. We preprocessed each posting by removing header data (for example, "from", "subject", "date", etc.) and lines quoted from previous messages being responded to (which start with a '>' character), performing stop-word removal and word stemming, and deleting the postings that became empty of text after these preprocessing phases.

**Table 1.** The topics from the "20-Newsgroups" used in the experiments.

| | | | |
|---|---|---|---|
| comp.graphics | alt.atheism | sci.crypt | misc.forsale |
| comp.sys.mac.hardware | talk.religion.misc | sci.electronics | rec.autos |
| comp.windows.x | talk.politics.guns | sci.med | rec.motorcycles |
| comp.os.ms-windows.misc | talk.politics.mideast | sci.space | rec.sport.hockey |
| comp.sys.ibm.pc.hardware | talk.politics.misc | soc.religion.christian | rec.sport.baseball |

In every independent run, we randomly selected 70% of the postings of each newsgroup to be used for training, and retained the remaining 30% for testing.

## 4.2 The Histograms/Features and Benchmarks Used

We first describe the process involved in the construction of the histograms and the extraction of the Quantile-based features.

Each document in the 20-Newsgroups dataset was preprocessed by word stemming using the Porter Stemmer algorithm and by a stopword removal phase. It was then converted to a BOW representation. The documents were then randomly assigned into training or testing sets.

The word-based histograms (please see Figure 2) were then computed for each word in each category by tallying the observed frequencies for that word in each training document in that category, where the area of each histogram was the total sum of all the columns. The CMQS points were determined as those points where the cumulative sum of each column was equal to the CMQS moments when normalized with the total area. For further clarification, we present an example of

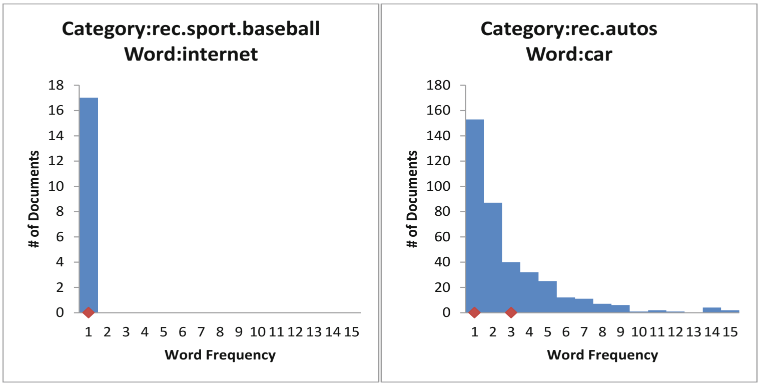


**Fig. 2.** The histograms and the $\frac{1}{3}$ and $\frac{2}{3}$ QS points for the two words "internet " and "car" from the categories "rec.sport.baseball" and "rec.autos".

two histograms[3] in Figure 2 below. The $\frac{1}{3}$ and $\frac{2}{3}$ QS points of each histogram are marked along their horizontal axes. In this case, the markings represent the word frequencies that encompass the $\frac{1}{3}$ and $\frac{2}{3}$ areas of the histograms respectively. The histogram on the left depicts a less significant word for its category while the histogram on the right depicts a more significant word for its category. Note that in both histograms the first CMQS point is located at unity. To help clarify the figure, we mention that for the word "internet" in "rec.sport.baseball", both the CMQS points lie at unity - i.e., they are on top of each other.

To compare the various methods used, we have developed three benchmarks for our system: A BOW classifier which involved the TFs and invoked the cosine similarity measure, a BOW classifier with the TFIDF features, and a Naïve-Bayes' classifier. Since they are well-established classifiers, their details are omitted – they are found in [8].

**The Metrics Used.** In every testing case, we used the respective data to train and test our classifier and each of the three benchmark schemes. For each newsgroup $i$, we counted the number of *True Positives* ($TP_i$) of postings correctly identified by a classifier as belonging to that group, the number of *False Negatives*, ($FN_i$) of postings that should have belonged in that group but were misidentified as belonging to another group, and the number of *False Positives* ($FP_i$) of postings that belonged to other groups but were misidentified as belonging to this one. The Precision $P_i$ is the proportion of postings assigned in group $i$ that are correctly identified, and the Recall $R_i$ is the proportion of postings belonging in the group that were correctly recognized. The $F$ score is an average of these two metrics for each group, and the *macro-F1* is the average of the $F$

[3] The documents used in this test were very short, which explains why the histograms are heavily skewed in favour of lower word frequencies.

scores over the all groups. All these are specified in Eq. (1).

$$P_i = \frac{TP_i}{TP_i + FP_i}; \quad R_i = \frac{TP_i}{TP_i + FN_i}; \quad F_i = \frac{2P_iR_i}{P_i + R_i}; \quad \textit{macro-F1} = \frac{1}{20}\sum_{i=1}^{20} F_i. \tag{1}$$

**Correlation Between the Classifiers.** Since the features and methods used in the classification are rather distinct, it would be a remarkable discovery if we could confirm that the results between the various classifiers are not correlated. Since the classifiers themselves yield binary results ('0' or '1' for incorrect or correct classifications), it is appropriate to compare classifiers $X$ and $Y$ by the "number" of times they yield *identical* decisions. In other words, a suitable metric for evaluating how any two classifiers $X$ and $Y$ yield identical results is given by Eq. (2) below:

$$\text{ClassifierSim}_{X,Y} = \frac{Pos_X Pos_Y + Neg_X Neg_Y}{Pos_X Pos_Y + Pos_X Neg_Y + Neg_X Pos_Y + Neg_X Neg_Y}, \tag{2}$$

where $Pos_X Pos_Y$ and $Neg_X Neg_Y$ are the count of cases where the classifiers $X$ and $Y$ both return identical decisions '1' or '0' respectively, and where '0' and '1' represent the events of a classifier classifying a document incorrectly or correctly respectively. Analogously, $Pos_X Neg_Y$ and $Neg_X Pos_Y$ are the counts of cases where $X$ returns '1' and $Y$ returns '0' and vice-versa respectively. Although this is a statistical measure of the relative similarities between the classifiers, we shall refer to this as their mutual "correlation".

## 5 Experimental Results

### 5.1 The Results Obtained: "Anti"-Bayesian TF Scheme

The experimental results that we have obtained for the "Anti"-Bayesian scheme that used only the TF criteria are briefly described below. We performed 100 tests, each one using a different random 70%/30% split of training and testing documents. We then evaluated the results of each classifier by computing the Precision, Recall, and $F$-score of each newsgroup, whence we computed the *macro-F1* value for each classifier over the 20-Newsgroups. The average results we obtained, over all 100 tests, are summarized in Table 2.

The results show that for *half* of the CMQS pairs, the "Anti"-Bayesian classifier performed as well as and sometimes even better than the BOW classifier.

Figure 3 displays the plots of the correlation between the different classifiers for the 100 classifications achieved, where in the case of the "Anti"-Bayesian scheme, the method used the TF features. The reader should observe the uncorrelated nature of the classifiers when the CMQS points are non-central [8].

**Table 2.** The *macro-F1* score results for the 100 classifications attempted and for the different methods. In the case of the "Anti"-Bayesian scheme, the method used the TF features.

| Classifier | CMQS Points | *macro-F1* Score |
|---|---|---|
| "Anti"-Bayesian | 1/2, 1/2 | 0.709 |
| | 1/3, 2/3 | 0.662 |
| | 1/4, 3/4 | 0.561 |
| | 1/5, 4/5 | 0.465 |
| | 2/5, 3/5 | 0.700 |
| | 1/6, 5/6 | 0.389 |
| | 1/7, 6/7 | 0.339 |
| | 2/7, 5/7 | 0.611 |
| | 3/7, 4/7 | 0.710 |
| | 1/8, 7/8 | 0.288 |
| | 3/8, 5/8 | 0.686 |
| | 1/9, 8/9 | 0.264 |
| | 2/9, 7/9 | 0.515 |
| | 4/9, 5/9 | 0.713 |
| | 1/10, 9/10 | 0.243 |
| | 3/10, 7/10 | 0.631 |
| BOW | | 0.604 |
| BOW-TFIDF | | 0.769 |
| Naïve-Bayes | | 0.780 |

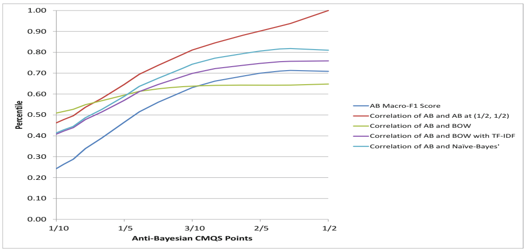


**Fig. 3.** Plots of the correlation between the different classifiers for the 100 classifications achieved. In the case of the "Anti"-Bayesian scheme, the method used the TF features.

### 5.2 The Results Obtained: "Anti"-Bayesian TFIDF Scheme

The results of the "Anti"-Bayesian scheme when it involves TFIDF features are shown in Table 3.

1. The results show that for *all* CMQS pairs, the "Anti"-Bayesian classifier performed much better than the traditional BOW classifier. For example, while the BOW had a *macro-F1* score of 0.604, the corresponding index for the CQMS pairs $\langle \frac{1}{3}, \frac{2}{3} \rangle$, was significantly higher, i.e., 0.747. Further, the *macro-F1* score indices for $\langle \frac{1}{4}, \frac{3}{4} \rangle$, $\langle \frac{3}{7}, \frac{4}{7} \rangle$ and $\langle \frac{4}{9}, \frac{5}{9} \rangle$ were consistently higher

**Table 3.** The *macro-F1* score results for the 100 classifications attempted and for the different methods. In the case of the "Anti"-Bayesian scheme, the method used the TFIDF features.

| Classifier | CMQS Points | *macro-F1* Score |
|---|---|---|
| "Anti"-Bayesian | 1/2, 1/2 | 0.742 |
| | 1/3, 2/3 | 0.747 |
| | 1/4, 3/4 | 0.746 |
| | 1/5, 4/5 | 0.742 |
| | 2/5, 3/5 | 0.745 |
| | 1/6, 5/6 | 0.736 |
| | 1/7, 6/7 | 0.729 |
| | 2/7, 5/7 | 0.747 |
| | 3/7, 4/7 | 0.744 |
| | 1/8, 7/8 | 0.720 |
| | 3/8, 5/8 | 0.746 |
| | 1/9, 8/9 | 0.712 |
| | 2/9, 7/9 | 0.745 |
| | 4/9, 5/9 | 0.744 |
| | 1/10, 9/10 | 0.705 |
| | 3/10, 7/10 | 0.748 |
| BOW | | 0.604 |
| BOW-TFIDF | | 0.769 |
| Naïve-Bayes | | 0.780 |

– 0.746, 0.744 and 0.744 respectively. This demonstrates the validity of our counter-intuitive paradigm – that we can truly get a remarkable accuracy even though we are characterizing the documents by the syntactic features of the points quite distant from the mean and more towards the extremities of the distributions.

2. In all the cases, the values of the Macro-*F1* index was only slightly less than the indices obtained using the BOW-TFIDF and the Naïve-Bayes approaches.

Figure 4 displays the plots of the correlation between the different classifiers for the 100 classifications achieved, where in the case of the "Anti"-Bayesian scheme, the method used the TFIDF features. The reader should again observe the uncorrelated nature of the classifiers for non-central CMQS points. This correlation increases as the feature points become closer to the mean/median.

To continue the analysis, it would be good to examine if the two "Anti"-Bayesian classifiers are relatively uncorrelated in and of themselves. Thus, if a particular pair of CMQS points yielded distinct classification decisions using the two schemes, and if they, all the same, yielded comparable accuracies, the potential of the paradigm is shown to be significantly more. This is precisely what we embark on achieving now – i.e., examining the correlation (or lack thereof) of the "Anti"-Bayesian TF and TFIDF schemes. This correlation is depicted graphically in Figure 5 whence the trends in the correlation with the increasing values of the CMQS points is clear.

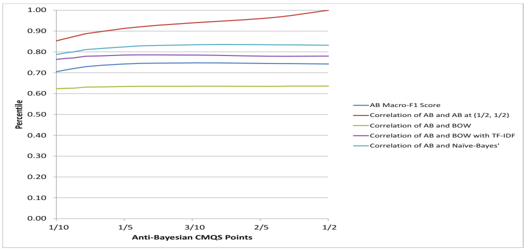


**Fig. 4.** Plots of the correlation between the different classifiers for the 100 classifications achieved. In the case of the "Anti"-Bayesian scheme, the method used the TFIDF features.

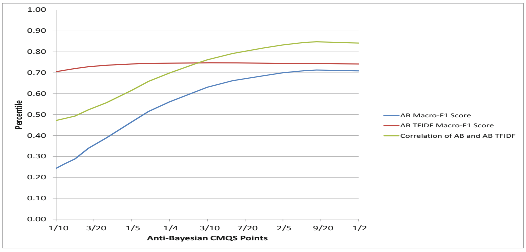


**Fig. 5.** The correlation between the two "Anti"-Bayesian classifiers for the 100 classifications when they utilized the TF and the TFIDF features respectively.

When the CMQS points are close to the mean or median, the correlation is quite high (for example, 0.842). This is not surprising at all, since in such cases, the "Anti"-Bayesian classifier reduces to become a Bayesian classifier. Also, when the CMQS points are far from the mean or median, the correlation is quite high (for example, 0.659 for the CMQS points $\langle \frac{2}{9}, \frac{7}{9} \rangle$). This is quite surprising because although both schemes are "Anti"-Bayesian in their philosophy, the lengths of the documents play a part in determining the decisions that they individually make because the IDF values account for document lengths.

The unabridged version of the paper [8] also describes how the various classifiers can be fused. This discussion is omitted here in the interest of space.

## 6 Conclusions

In this paper we have considered the problem of Text Classification (TC), which is a problem that has been studied for decades. From the perspective of classification, problems in TC are particularly fascinating because while the feature extraction process involves *syntactic or semantic* indicators, the classification uses the principles of *statistical* Pattern Recognition (PR). The state-of-the-art in TC uses these statistical features in conjunction with the well-established methods such as the Bayesian, the Naïve Bayesian, the SVM etc. Recent research has advanced the field of PR by working with the Quantile Statistics (QS) of the features. The resultant scheme called Classification by Moments of Quantile Statistics (CMQS) is essentially "Anti"-Bayesian in its *modus operandus*, and advantageously works with information latent in "outliers" (i.e., those distant from the mean) of the distributions. Our goal in this paper was to demonstrate the power and potential of CMQS to work within the *very* high-dimensional TC-related vector spaces and their "non-central" quantiles. To investigate this, we considered the cases when the "Anti"-Bayesian methodology used both the TD and the TFIDF criteria.

Our PR solution for $C$ categories involved $C-1$ pairwise CMQS classifiers. By a rigorous testing on the well-acclaimed data set involving the 20-Newsgroups corpus, we demonstrated that the CMQS-based TC attains accuracy that is comparable to and sometimes even better than the BOW-based classifier, even though it essentially uses the information found only in the "non-central" quantiles. The accuracies obtained are comparable to those provided by the BOW-TFIDF and the Naïve Bayes classifier too!

Our results also show that the results we have obtained are often uncorrelated with the established ones, thus yielding the potential of fusing the results of a CMQS-based methodology with those obtained from a more traditional scheme.

## References


1. Alahmadi, A., Joorabchi, A., Mahdi, A.E.: A new text representation scheme combining bag-of-words and bag-of-concepts approaches for automatic text classification. In: Proceedings of the 7th IEEE GCC Conference and Exhibition, Doha, Qatar, pp. 108–113, November 2014
2. Debole, F., Sebastiani, F.: Supervised term weighting for automated text categorization. In: Proceedings of the 18th ACM Symposium on Applied Computing, Melbourne, USA, pp. 784–788, March 2003
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. A Wiley Interscience Publication (2006)
4. Dumoulin, J.: Smoothing of n-gram language models of human chats. In: Proceedings of the Joint 6th International Conference on Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), Kobe, Japan, pp. 1–4, November 2012
5. Lu, L., Liu, Y.-S.: Research of english text classification methods based on semantic meaning. In: Proceedings of the ITI 3rd International Conference on Information and Communications Technology, Cairo, Egypt, pp. 689–700, December 2005
6. Madsen, R.E., Sigurdsson, S., Hansen, L.K., Larsen, J.: Pruning the vocabulary for better context recognition. In: Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, vol. 2, pp. 483–488, August 2004
7. Ning, Y., Zhu, T., Wang, Y.: Affective-word based chinese text sentiment classification. In: Proceedings of the 5th International Conference on Pervasive Computing and Applications (ICPCA), Maribor, Slovenia, pp. 111–115, December 2010

8. Oommen, B.J., Khoury, R., Schmidt, A.: Text Classification Using "Anti"-Bayesian Quantile Statistics-based Classifiers. Unabridged version of this paper. Submitted for publication
9. Oommen, B.J., Thomas, A.: Optimal Order Statistics-based "Anti-Bayesian" Parametric Pattern Classification for the Exponential Family. Pattern Recognition **47**, 40–55 (2014)
10. Qiang, G.: An effective algorithm for improving the performance of naïve bayes for text classification. In: Proceedings of the Second International Conference on Computer Research and Development, Malaysia, pp. 699–701, May 2010
11. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. Mc-Graw Hill Book Company, New York (1983)
12. Salton, G., Yang, C.S., Yu, C.: Term weighting approaches in automatic text retrieval. Technical Report, Ithaca, NY, USA (1987)
13. Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys **34**, 1–47 (2002)
14. Thomas, A., Oommen, B.J.: The Fundamental Theory of Optimal "Anti-Bayesian" Parametric Pattern Classification Using Order Statistics Criteria. Pattern Recognition, 376–388 2013
15. Thomas, A., Oommen, B.J.: Order Statistics-based Parametric Classification for Multi-dimensional Distributions. Pattern Recognition, 3472–3482 (2013)
16. Thomas, A., Oommen, B.J.: Corrigendum to Three Papers that deal with "Anti"-Bayesian Pattern Recognition. Pattern Recognition, 2301–2302 (2014)
17. Wu, G., Liu, K.: Research on text classification algorithm by combining statistical and ontology methods. In: Proceedings of the International Conference on Computational Intelligence and Software Engineering, Wuhan, China, pp. 1–4, December 2009